Final paper for Course T-61.184: Survey Project - Segment-based Speech Recognition

Petri Korhonen

Helsinki University of Technology
petri@acoustics.hut.fi

Abstract

Most speech recognition systems take as an input a set of features computed at fixed frame rate from short-time windows of speech signal. An alternative to this framework is to perform prior to the recognition phase some acoustical analysis to get some explicit segmentation of the speech signal. The segments obtained this way are of variable lengths, and different features can be used for different segments. The segmental framework allows a richer set of acoustic phonetic features than can be incorporated into conventional frame-based representations. In this survey project this framework will be studied in detail. This course paper is divided in two parts: first a general framework of segment based ASR as presented by Mari Ostendorf et al. [1], the latter part of this paper presents the details of the SUMMIT segment-based speech recognition system developed at MIT.

1. Introduction

In the last two decades the automatic speech recognition has dramatically improved. The use of mathematically rigerous Hidden-Markov Models (HMMs) has in part contributed to this. The acoustic models in HMM based systems model the temporal sequence of feature vectors computed at fixed frame rate, most commonly 10ms/frame. The duration of the typical phoneme can vary from 20ms over to 200ms, thus the number of fixed-rate feature vectors within the same phonetic segment is usually much greater than one. Since in the speech production the articulators move rather slowly, the feature vectors within the same phonetic segment are usually highly correlated. However, HMMs have an inherent conditional independence assumption on the observation feature vectors. This in turn means that the fixed frame-rate feature vector used is HMM-based recognizers fundamentally limits the range of acoustic models that can be explored for encoding acoustic-phonetic information [2].

In segment-based speech recognition system acoustic models model a sequence of feature vectors computed at time intervals that are not necessarily equal. The segment features are computed from a portion of the speech waveform belonging to a hypothesized phonetic segment. Briefly, the advantages of using segments are listed in [1]:

- Segments provide a better framework for modelling statistical dependence among spectra in nearby frames.
- Average segment duration can be much greater than the frame duration typically used in ASR systems leading to computational savings.
- Segment boundaries usually occur at points of large spectral change. There is evidence that such points may be particularly rich in phonetic information about the identity of certain consonants. Thus, measurements made near these points may be found to be useful.

2. Segmental model vs. HMM

In [1] Ostendorf et. al. attempted to bring together a variety of work done under a common framework in order to make it easier for different researchers to benefit from the successes of others. The research on the new techniques (namely segment-based modeling) has tended to proceed in isolated pockets. Here the basic idea of segment models (SM) are introduced following their paper.

The statistical approach for speech recognition involves finding the underlying sequence of labels $a_1^N = \{a_1, \ldots, a_N\}$ that is most likely given the sequence of T D-dimentional feature vectors $y_1^T = \{y_1, \ldots, y_T\}$. In mathematical terms this can be written as

$$\hat{a}_1^N = \arg\max_{N,a_1^N} \mathrm{p}(a_1^N|y_1^T) = \arg\max_{N,a_1^N} \mathrm{p}(a_1^N)\mathrm{p}(y_1^N|a_1^T) \quad (1)$$

where a_i corresponds for example to phone, and the recognized phone sequence is constrained to pronunciations in a lexicon. a does not need to be a phone label, but can be a longer unit such as biphone, triphone, or some

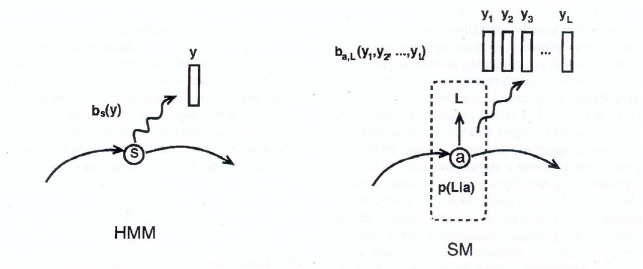


Figure 1: Hidden Markov model (HMM) and Segmental Model (SM) illustrated as generative processes. (From [1])

other automatically learned unit such as syllable. Equation 1 can be interpreted as consisting of a language model p(a_1^N), and an acoustic model p($y_1^T|a_1^N$). The fundamental difference in HMM-based ASR systems and segment-based systems is namely in the acoustic model. In HMMs the the fundamental observation distribution is at frame level, where as in segment modeling the fundamental distribution model $b_{a,l}(y_1^l)$ = p($y_1^l|a,l$) represents a segment $y_1^l = [y_1, \ldots, y_l]$, where l is a random variable, and $a \in A$, where A is the set of segment labels. The difference of these models is illustrated in Figure 1 from the perspective of generative models. The "segment" can consist from phoneme to larger unit, and it does not affect the probabilistic formalism presented here. In both HMMs and SMs the discrete state sequence s_i^T, and $(a,l)_1^N$ respectively is typically modeled as a markov chain. In both HMMs and SMs there exists several options for modeling the distribution of observations, including dicrete distrtibutions, full or diagonal covariance Gaussian densities, Gaussian mixtures and Laplacian distributions. With SMs there are in fact even more options because of the large number degrees of freedom.

2.0.1. General Modeling Framework

Segment model in general form provides a joint model for observation sequences of random length $y_1^l = [y_1, \ldots, y_l]$, generated by unit a with the density

$$\mathrm{p}(y_1, \ldots, y_l|a) = \mathrm{p}(y_1, \ldots, y_l|l, a)\mathrm{p}(l|a) = b_{a,l}(y_1^l)\mathrm{p}(l|a) \tag{2}$$

From the above equation we can see that the segment model for label a is characterized by 1) a duration distribution $p(l|a)$, and 2) a family of output densities $b_{a,l}(y_1^l); l \in L$ that describes observation sequences of different lengths. Also a Markov assumption for sequences of a_i is made either implicitly or explicitly by embedding phone segments in a word pronunciation network or other probabilistic finite-state network.

The simplest distribution assumption for a segment model uses a single output distribution and assumes independence between successive observed frames, and that they are identically distributed within given segment boundaries. This can be seen as a one-state HMM with an explicit duration model

$$b_{a,l}(y_1^l) = \prod_{i=1}^{l} p(y_i|a) \tag{3}$$

This is called hidden "semi-Markov" model. Introduction of the explicit state duration model adds complexity of hypothesizing segmentations in recognition and training.

Next step of increasing the complexity is to use multiple distribution regions $r = 1, \ldots, R$ still assuming that observations are conditionally independent given the segment length, and we have

$$b_{a,l}(y_1^l) = \prod_{i=1}^{l} p(y_i|a, r_i) \tag{4}$$

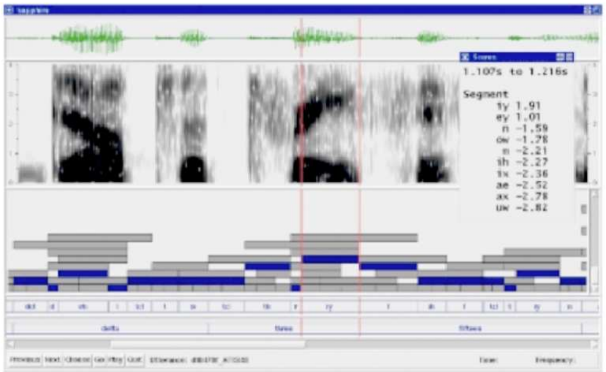
where the specific distribution used for y_i depends on the region r_i. This can be reduced to represent HMM with complex topology. From here the segment model can be generalized further in a variety of ways.

The segment duration distribution (in Eq. 2) can be either parametric or non-parametric. For phone-sized unit any reasonable distribution assumption works well empirically, because the contribution of the duration model is small relative to the segment observation probability. The family of output densities $b_{a,l}(y_1^l); l \in L$ represents l-length trajectories in vector space $(y_i \in \Re^d)$ with a sequence of distributions that can be thought as a dividing the segment into separate regions in time. Observations can correlate within and across regions. Nevertheless, the distribution parameters are time invariant within a region. Segment distribution region is in some sence analogues to an HMM state. The collection of distribution mappings $T_l(i); i = 1, \ldots, l; l \in L$ associate each frame of the observation vector y_i in the the observation sequence with one of the model regions. The mapping and the region dependent distributions provide a means of specifying $b_{a,l}(y_1^l)$ for a large range l with a small number of parameters.

The mapping T_l is a key component for specifying the distribution family. T_l can be deterministic or dynamic.

2.0.2. Recognition Algorithms

The recognition algorithms for the segment models is similar to that used for HMMs. In HMMs the standard recognition solution involves finding the most likely state sequence via viterbi decoding and then mapping the state sequence to the appropriate word sequence. For segment



```
1.107s to 1.216s

Segment
      iy 1.91
      ey 1.01
       n -1.59
      ow -1.78
       m -2.21
      ih -2.27
      ix -2.36
      ae -2.52
      ax -2.78
      uw -2.82
```

Figure 2: Multi level segmentation of an utterance. From up, speech waveform, spectrogram, hypothesised phonetic segmentation, best-scoring phonetic sequence, best-scoring word sequence. (From [3])

models, the solution is analogous, but in this case, the state includes both the segment label and duration, and recognition involves finding

$$(\hat{N}, \hat{a}_1^{\hat{N}}) = \underset{N,a_1^N}{\arg\max}\{\underset{l_1^N}{max}\, \mathrm{p}(y_1^T|l_1^N, a_1^N)\mathrm{p}(l_1^N|a_1^N)\mathrm{p}(a_1^N)\} \quad (5)$$

using a dynamic programming algorithm and then mapping the sequence to the appropriate word sequence, and then mapping the segment label sequence to the appropriate word sequence. The key difference between the SM and HMM search algoritrhms is the explicit evaluation of different segmentations. Segmentation adds an extra dimension to the dynamic programming search. Reducing the search space is a key to reducing number of segment evalutations. This is done by introducing a multi-level segmentation of the signal prior to segment evaluations. An example of a multilevel segmentation (dendogram) is shown in Figure 2.

2.0.3. Parameter Estimation Algotithms

The hidden state component in acoustic modeling require some kind of iterative algoritm for maximum likelihood (ML) estimation of parameters. Ostendorf et. al. represent two algoritms, expectation-maximization (EM) and ”Viterbi training”, in generalized form that can be used in both segmental and HMMs.

2.1. Models of Feature Dynamics

HMM can be seen as a special case of a segment model, thus segment model is capable of acheieving at least the same level of performance as an HMM. However, segment model allows more more general families of distribution than with HMM, especially distributions that model feature dynamics. There exists many alternatives for distribution assuptions that can represent feature dynamics such as “contrained mean”, Gauss-Markov, more general linear models, and segmental mixture models. Each of these are described in detail in [1].

2.2. SUMMIT Speech Recognition System

The segment-based framework for speech recognition can be used in number of ways, and not all the systems follow the same formalism. In this section SUMMIT speech recognition system is introduced. SUMMIT is probably the largest and most well known segment based ASR system developed to this date. SUMMIT does not follow exactly the formalism outlined above, but is presented here as an example of a real segment based ASR system.

SUMMIT system uses segment-based framework for its acoustic-phonetic representation of the speech signal. Acoustic and/or probabilistic landmarks form a basis for a phonetic segment network, or graph as in Figure 2. Feature vectors are extracted both over hypothesized segments and at their boundaries to be used in phonetic analysis; the system uses two types of acoustic models: *segment* models and *boundary* models. The obtained feature space for the speech signal takes the form of an acoustic-phonetic network, whereby different paths through the network are associated with different sets of feature vectors. This is quite different from prevailing approaches which employ a temporal sequence of observation vectors, which typically contain short-time spectral information. Segment models can be be context-independent or context-dependent.

One requirement of segment based ASR system is to explicitly hypothesize segment boundaries i.e. segment start and end times. It would be computationally expensive to model and search all possible segments. In SUMMIT system two types of segmentation schemes have been used. One is based on finding the instances of most acoustic variability in signal. This method does not provide a comprehensive segmentation, and usually much greater number of segments are created than there are phonemes in an utterance. The other segmentation method is ”segmentation by recognition”, whereby a segment network is obtained in process of running a first pass recognizer with a suitable search. This could be for example forward pass Viterbi search of all possible segmentations of set of frames. ”Segmentation by recognition” has several desirable charactereistics. It is flexible, because it can use any first pass recognition strategy; it does not rule out the use of local acoustic measures, or combination of context-dependent acoustic measures and language models. Also ”segmentation by recognition” is accurate, and minimizes the number of deletion and insertion errors. Finally the first pass recognition result can be used as such in subsequent segment-based recognition taking advantage of different

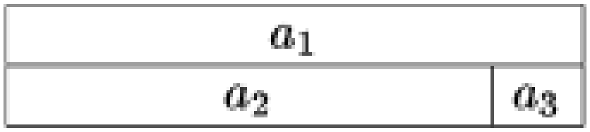


Figure 3: A hypothetical segment network that contains three features, a_1, a_2, and a_3, and two segmentations. (From [4])

recognition strategies.

In frame based ASR systems the set of acoustic features A is a temporal sequence of vectors. Each segmentation S of the speech signal accounts for all frames and therefore all A, making it efficient to compute $P(A|W)$. For segment-based recognition, A is a temporal network of features and each segmentation S accounts on a subset of all segments and and therefore only a subset, A_S of A. This makes it impossible to compare different paths through the network with out some normalization. In order for a path through S to account for all A, it must also account for $A_{\bar{S}}$, $A = A_S \cup A_{\bar{S}}$. Figure 3 illustrates this. For this hypothetical network there A contains three features a_1, a_2 and a_3, and two segmentations. When examining a path through top segmentation, we must account for both $A_{S_{top}}$, containing a_1, and $A_{\bar{S}_{top}}$ containing a_2 and a_3. This leads for each segmentation, S:

$$P(A|W) = P(A_S A_{\bar{S}}|W) \tag{6}$$

this means that the entire segmentation network must be processed once for each segmentation, and because the number of possible segmentations grows exponentially as the number of segments, such processing is conputationally daunting. To overcome this SUMMIT system has two strategies: "not" modeling, and "near-miss" modeling.

2.2.1. "Not" modeling

"Not" modeling is an algorithm that efficiently computes $P(A_S A_{\bar{S}}|W)$ for segment based regognition using and additional nonlexical "not" model, $\bar{\omega}$, to account for all segments that are not in a segmentation and therefore $A_{\bar{S}}$. Assuming independence between A_S and $A_{\bar{S}}$, not model $\bar{\omega}$ can be used to normalize each segmentation, S, to implicitly account for all segments:

$$P(A_S A_{\bar{S}}|W) = P(A_S|W)P(A_{\bar{S}}|\bar{\omega})\frac{P(A_S|\bar{\omega})}{P(A_S|\bar{\omega})} = K\frac{P(A_S|W)}{P(A_S|\bar{\omega})} \tag{7}$$

K being a constant for all segmentations. Rather than scoring A_S against lexical models and $A_{\bar{S}}$ against "not" model it is sufficient to score A_S against all models, both lexical and "not" models.

2.2.2. Near-miss modeling

"Not" modeling is efficient, but it has shortcomings. In "not" modeling all segments that are not in the segmentation S are mapped into a single class, even though many segments that are not in a segmentation are as distinct as the segments that are in the segmentation. The other suggested method in SUMMIT system is a generalized idea of ”not” modeling called ”near-miss” modeling. In this algorithm multiple nonlexical classes are used to model segments that are not in the segmentation, but are ”near-misses” of those which are.

To efficiently compute $P(A_S A_{\bar{S}}|W)$ for near-miss modeling, SUMMIT introduces a search algorithm which associates each segment with a near-miss subset drawn from all other segments in the network. The near-miss subsets are drawn such that for each segmentation, S, the near-miss subsets of the segments in S are mutually exclusive and their union, $\bar{A}$, is $A_{\bar{S}}$. If independence between A_S and $\bar{A}_S$ is assumed, for each segmentation S we can compute

$$P(A_S A_{\bar{S}}|W) = P(A_S|W)P(\bar{A}_S|\bar{W}) \tag{8}$$

$\bar{W}$ being the nonlexical models associated the the segments that are not in S. It has been shown that there exist such near-miss subsets for any segment network [4]

Near-miss modeling maintains the integrity of the probabilistic framework, allowing the use of efficient search strategies, such as viterbi, to enforce context accross the entire segment network. It has the potential for more sophisticated modeling than the anti-phone method since it is general and allows the segments in the network to be modeled in any manner. [3]

2.2.3. Landmark modeling

In addition to modeling segments, in SUMMIT system additional information about segment boundaries, or landmarks, is computed. [3] These landmarks Z adds a new dimensionality to observation space A. If we denote segments with X, and near-misses with Y, we need to estimate the probability $P(X, Y, Z|S, U)$. It is reasonable to assume independence between the segments and the landmarks changing the computation of probability to

$$P(X, Y, Z|S, U) = P(X, Y|S, U)P(Z|S, U) \tag{9}$$

Landmark models are computed at the positions all possible segment boundaries of the segment network, thus a particular segmentation will assign some of the landmarks to transitions between lexical units, while the remainder of landmarks will be considered to occur internal to a unit. Every segmentation accounts for all of the landmark observations Z making it unnecessary to employ any normalisation criterion such as was in case of graph-based observations. It is reasonable to assume conditional independence between m individual observations in Z given U, and the probability $P(Z|S,U)$ can be written as

$$P(Z|S,U) = \prod_{i=1}^{m} P(z_i|S,U) \tag{10}$$

where z_i is the observation extracted are the ith landmark. Landmarks are comparable to frames, because they are are sequental in nature, but they occur at uneven intervals in time, and they occur much less frequently, making the conditional independence assumption between independent landmarks more reasonable than that made in frame-based methods. Overall the segmental measurements can be interpreted to quantify within-phone dynamics, whereas landmark measurements quantify transitions between phones.

2.2.4. *Decoding Implementation in SUMMIT*

Since the features in segment based system does not follow each other in sequential manner decodong algorithms has to be modified to overcome this. In SUMMIT system recognition is done via a modified Viterbi algoritm that can be viewed as finding the best path through two graphs; a conventional pronunciation graph representing all possible word sequences and their associated pronunciations, and an acoustic phonetic graph representing all possible segmentations of spoken utterance. This can be formulated as finding a best path through a graph $A \circ U$ where A is the acoustic-phonetic graph, and U is the pronunciation graph composed of conversion of context-dependent to context-independent lexical labels, phonological rules, mappings of pronunciations to words in the lexicon, and language model. This is done with finite state transducers [5].

Training is done by Viterbi-style training whereby forced alignments of ortographic transcriptions are computed to create reference phonetic transcriptions that are used to train acoustic-phonetic models. Aforementioned context dependent landmark models are trained based on the forced alignment segmentations. Between-phone landmarks are modeled as transitions, and intra-segmental landmarks are modeled as internal boundaries for a given phone. Anti-phone or near-miss models are trained on all segments outside the forced alignment segmentation through a segment graph.

2.2.5. *Performance of the System*

It is not the scope of this survey project to go into specifics of any system, but the aim was to research the framework itself. However, here some recognition results achieved with the segment based system (SUMMIT) is presented to show how far this framework can get us. It also gives some idea what the features used in a practical system might be. The SUMMIT system has been tested on several domains over the years [6] [7], but here recognition results for phonetic recognition are presented as reported by Glass et. al. in [3].

Feature extraction is based on averages and derivatives of 12 MFCCs and PLP cepstral coefficients plus energy and duration. Acoustic models are based on mixtures of diagonal Gaussians. Prior to modeling the acoustic feature space is whitened with a global rotation, which transforms the pooled within-class covariance of the training data to identity matrix to get uncorrelated training data, which has unity variance across all dimensions. Also this technique reduces the dimensionality of the feature vector itself.

All the test were based on widely used TIMIT acoustic-phonetic corpus. Glass et al. reported phonetic recognition error which included substitution, deletion, and insertion errors. The language model used was phone bigram based on the training data. A single parameter controlled the trade-off between insertions and deletions.

Combination of different configurations of segment-based/landmark and anti-phone/near-miss models were tested. A probabilistic segmentation, which produced segment graphs with a density of approximately 60 segments/s compared to the 13 segments/s found in an average TIMIT phonetic transcription, was used. Best results were achieved with configuration where landmark models were used in tandem with segment models and anti-phone models were used. In this case the reported phone error rate was 24.4%. In the Table 1 the best result is compared with the best results reported in the literature. There are differences regarding the complexity of the acoustic and language models, making the comparison difficult, but Glass et. al. believe that the results are a strong indication for the viability of a segment-based approach.

Table 1: Reported phonetic recognition error rates on the TIMIT core test set (from [3])

Method	Phone error rate (%)
Triphone CDHMM	27.1
Recurrent Neural Network	26.1
Bayesian Triphone HMM	25.6
Near-miss, probilistic segmentation	25.5
Anti-phone, Heterogeneous classifiers	24.4

A nice property of the SUMMIT system is its fast training cycle; the distributed computation allows one complete acoustic training iteration of 100 hours of speech in 5-6 hours. Also since the converge is fast the whole ASR system can be trained in less than a day. In the HMM based systems the training can take several weeks.

3. Conclusions

In this course paper a segment based speech recognition framework was reviewed. First part of the paper concentrated on the work done by Mari Ostendorf et. al.. The latter part of the paper described the most well known segment-based ASR system called SUMMIT.

Segment-based framework is not strictly defined, and development of segment-based speech recognition ideas has happened in more or less isolation. Paper by Ostendorf et. al. tried to define a common language for researchers in different research lab to understand the advances made by others in the field. This is by no means an easy task, since there are no standard tools or algorithms that people use exclusively.

The main advantages of segment-based modeling of speech signal are: many alternatives for representing a family of distributions (allowing trajectory modeling and/or correlation modeling), and possibility for models for intra-segmental timing. The main disadvantage is the need of explicit segmentation. There have been numerous of attempts to create explicit segmentation, but so far use of segment networks (dendograms) has been the answer to overcome this problem.

Frame based systems have provided the best results for speech recognition in the recent years. The main question raised by those developing the segental speech recognition is, is the good performance of HMM based systems because of the use of HMM or is it because of the highly sophisticated algorithms developed by numerous of scientist developing ASR systems using the frame-based framework. There has not been many ambitious attempts to develop a segment-based ASR systems, so comparison between frame-based and segment-based state-of-the-art ASR systems is not straighforward.

4. References

[1] Jeffrey N. Marcus, “Phonetic recognition in a segment-based HMM”, Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, Volume: 2, 27-30 April 1993, Pages:479 - 482 vol.2

[1] Mari Ostendorf and Vassilios V. Digalakis and Owen A. Kimball, “From HMM’s to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition”, IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 5, September 1996

[2] Han Shu and I. Lee Hetherington and James Glass, “Baum-Welch Training for Segment-Based Speech Recognition”, Automatic Speech Recognition and Understanding, 2003. ASRU ’03. 2003 IEEE Workshop on , 30 Nov.-3 Dec. 2003, Pages:43 - 48

[3] James R. Glass, “A probabilistic framework for segment-based speech recognition”, Computer Speech and Language 17 (2003) 137-152.

[4] Jane W. Chang and James R. Glass, “Segmentation and Modeling in Segment-Based Recognition”, Proc. Eurospeech ’97, Rhodes, Greece, 1199–1202, 1997,

[5] Glass, J., Hazen, T., Herrington, L., “Real-time telephone-based speech recognition in Jupiter domain” 1999. Proc. ICASSP Phoenix, AZ, pages: 61-64, March.

[6] Chang, J. “Near-miss modeling: a segment-based approach to speech recognition” Ph.D. thesis, EECS, MIT.

[7] Livescu, K., Glass, J., “Segment-based recognition on the PhoneBook task: initial results and observations on duration modeling” Proc. Eurospeech Aalborg, Denmark, September, pp. 1437-1440.